MORSE THEORY IN TOPOLOGICAL DATA ANALYSIS

HENRY ADAMS, ATANAS ATANASOV, AND GUNNAR CARLSSON

Abstract. We introduce a method for analyzing high-dimensional data. Our approach is inspired by Morse theory and uses the nudged elastic band method from computational chemistry. As output, we produce an increasing sequence of cell complexes modeling the dense regions of the data. We test the method on several data sets and obtain small cell complexes revealing informative topological structure.

Contents

1. Introduction

The analysis of large sets of high-dimensional data is a fundamental problem for all branches of science and engineering. Regression analysis can be used effectively when the data has a linear or low degree polynomial structure based on a choice of model for the data. However, it is often the case that the data is genuinely nonlinear and that there isn't an obvious choice for how to model it. The purpose of topological data analysis [Carlsson 2009] is to provide methods which produce simple combinatorial representations of the data.

In this paper we construct a cell complex representation in which the cell structure depends on the density of the points in the data set. We adapt the mathematical formalism of *Morse theory*, which in its idealized form constructs a cell decomposition of a manifold using sublevel sets of a function on the manifold (called the "Morse function"), to the setting of point clouds, i.e. finite sets of points in Euclidean spaces. The specific features of our approach are as follows.

- We use a density estimator as our analogue of the Morse function. One could also use other intrinsic functions on the geometry, such as notions of data depth, to obtain other compact representations.

- In order to sample cells from the analogue of the Morse skeleton of the Morse complex, we adapt the nudged elastic band method (NEB) from computational chemistry [Jónsson et al. 1998; Henkelman and Jónsson 2000] to our situation. NEB has been used to study high-dimensional conformation spaces of complicated molecules, and has typically used an energy function as the analogue of the Morse function.
- We produce an increasing sequence of cell complex models. In accordance with the idea of topological persistence, this output gives a more accurate representation of the data than the choice of any single complex.
- For the moment, we construct only the one-dimensional skeleton of the cell complex. Producing higher-order cells will require more difficult mathematics, since the minimization problems involved in the construction of such cells are challenging, and are related to minimal surface problems in geometric analysis.

In studying data sets computationally, one finds that outliers will generally obscure topological features. One approach for dealing with outliers is threshholding by density, i.e. studying superlevel sets of a function that estimates density. This is the approach taken by Carlsson et al. [2008] and Adams and Carlsson [2009], for example. One difficulty is that the superlevel sets of density frequently require large numbers of points to represent them, since they are "codimension zero" subsets. Consider Figure 1, which contains a standard Morse theoretic picture: a sublevel set of a torus in $\mathbb{R}^3$ with Morse function given by the z-coordinate. As is standard in Morse theory, the sublevel set is homotopy equivalent to the Morse skeleton, which is the dotted loop in Figure 1. Note that in order to achieve an accurate representation of the homotopy type, sampling from the entire sublevel set will require many more points than sampling from the Morse skeleton. The goal of this paper is to demonstrate that one can effectively sample cells from the Morse skeleton of a density function on the data set, thereby obtaining a more economical representation of the topology of the data set.

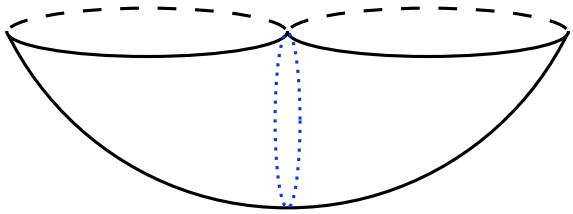

FIGURE 1. A sublevel set of a torus in $\mathbb{R}^3$ with Morse function given by the z-coordinate. The dotted loop is the Morse skeleton.

We demonstrate how the method works on several nonlinear test data sets, including sets arising in social networks, in image processing, and in microarray analysis. The results assist in our qualitative understanding of the data, and we discuss how this qualitative understanding may then be used. First, while it is common to cluster social network data, there is more understanding to be gained about the structure within a single cluster or about the relationships between different clusters. As illustrated with the social network data, our method can recover extra information. Second, the qualitative features we identify in image processing data have applications in compression and texture analysis. Third, microarray analysis has been used to identify genes which are part of the cell cycle [Spellman et al. 1998; Alter et al. 2000; Whitfield et al. 2002]. The expressions for these genes are recurrent but

not periodic: the amplitude and period of the expressions vary with time. One would like methods for detecting recurrent data which are robust to such changes. More generally, one may want to be able to recognize recurrent data in which the sampling times are irregular, spaced further apart than the period, or unknown. Our method offers tools in this direction.

We describe related work in Section 2; we provide background material on cell complexes, Morse theory, and NEB in Section 3; we describe our method in Section 4; we illustrate the results on test data sets in Section 5, and we conclude in Section 6.

2. Related work

Morse theory has previously been adapted to discrete and applied settings. Forman [2002] studies discrete Morse functions that assign a single value to each cell in a complex. Edelsbrunner et al. [2003] study piecewise linear Morse functions defined on triangulated manifolds, and Gyulassy et al. [2007] use a similar framework to analyze scalar functions arising in data analysis. By contrast, we will work with point clouds instead of triangulated spaces.

Given a point cloud data set drawn from an unknown probability density function, one may cluster the data by approximating the basins of attraction of the modes of density [Wishart 1969; Hartigan 1975, pp. 205]. The goal of Stuetzle and Nugent [2010] is to recover the full cluster tree of the density function, which describes how the connected components of the superlevel sets merge as the density threshold value decreases. Regardless of its topology, a connected component of a superlevel set is always represented in the cluster tree by a single point. There is more information to be gained by considering the topology of the superlevel sets. For example, some superlevel sets of the density function in Figure 2 have circular topology, but this is not apparent in the cluster tree.

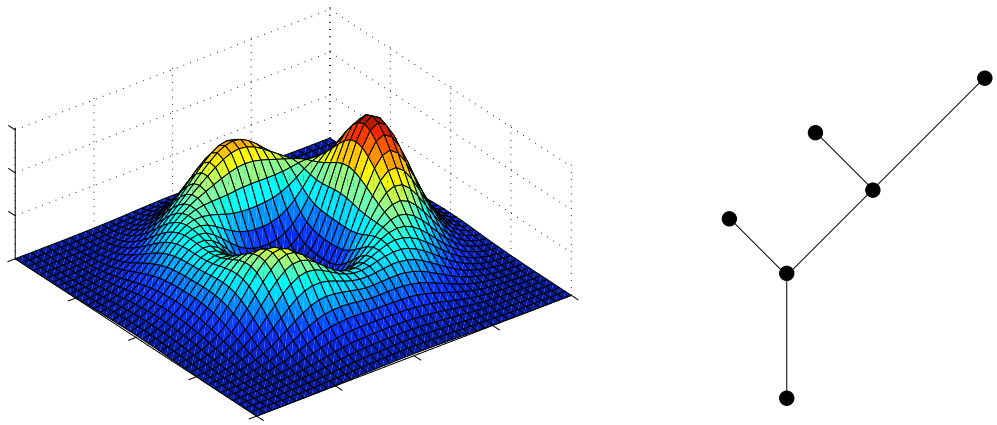

Figure 2. A probability density function on the left and the cluster tree for its superlevel sets on the right. Qualitatively, the function has a circular base and three bumps of varying heights. For a certain choice of density threshold the superlevelset of the function has the topology of a circle, but this is not reflected in the cluster tree.

In topological data analysis one often models a superlevel set of density with a simplicial complex, such as the Čech, Vietoris–Rips, Alpha, or witness complexes [Edelsbrunner and Mücke 1994; de Silva and Carlsson 2004]. These simplicial complexes alone are not useful descriptors of a superlevel set: they typically contain thousands of simplices and hence

are too large to interpret by hand. However, each of the Čech, Vietoris–Rips, Alpha, and witness complexes has a connectivity parameter that one can vary to produce an increasing sequence of simplicial complexes. One can compute the persistent homology of this increasing sequence to approximate the homology groups of the superlevel set [Edelsbrunner et al. 2002; Zomorodian and Carlsson 2005; Edelsbrunner and Harer 2010].

Though homology is a useful signature for describing the geometry of a superlevel set, in general there are several possible models with given homology groups, and determining which model best fits the data is a supervised step. One may use our Morse theoretic method to identify models which match not only the homology groups but also the geometry of the data, and so our method plays a complementary role to persistent homology. For example, Carlsson et al. [2008] use persistent homology to identify an image processing data set with $\operatorname{rank}(H_0) = 1$ and $\operatorname{rank}(H_1) = 5$, where H_i denotes the i-th homology group. There are many spaces satisfying these homological constraints, including a wedge sum of five circles. Carlsson et al. propose a particular space: a model containing three circles with four intersection points (Figure 7). In Section 5.2 we use our method to obtain this three circle model directly. Our model contains only 12 cells, much fewer than the number of simplices needed in a witness complex reconstruction.

Whereas Carlsson et al. [2008] and Adams and Carlsson [2009] study only one superlevel set of density at a time, the approach of Chazal et al. [2011] is similar to ours in that they study multiple superlevel sets with varying threshold values simultaneously.

3. Background

We briefly introduce three topics: CW complexes, Morse theory, and the nudged elastic band method.

3.1. **CW complexes.** A CW complex is a type of cell complex. For k a nonnegative integer, a k-cell is the closed ball $\{y \in \mathbb{R}^k \mid \|y\| \leq 1\}$ of dimension k. So a 0-cell is a point, a 1-cell is a line segment, a 2-cell is a disk, etc. A CW complex W is a topological space formed by the following inductive procedure. The 0-skeleton $W^{(0)}$ of W is a set of 0-cells. The 1-skeleton $W^{(1)}$ is formed by gluing the endpoints of 1-cells to the 0-skeleton, and can be thought of as a graph. Inductively, we form the k-skeleton $W^{(k)}$ by gluing the boundaries of k-cells to the $(k-1)$-skeleton $W^{(k-1)}$. If W is a finite-dimensional CW complex, then this process terminates and we have $W = W^{(k)}$ for some k. See Figure 3 for an example and Hatcher [2002] for further details.

3.2. **Morse theory.** The following introduction to Morse theory is informal; see Milnor [1965] for a thorough treatment. Suppose M is a compact manifold of dimension d and $f\colon M \to \mathbb{R}$ is a smooth function with non-degenerate critical points $m_1, \ldots, m_k \in M$ that satisfy

$$a_0 < f(m_1) < a_1 < f(m_2) < \cdots < a_{k-1} < f(m_k) < a_k.$$

The index λ_i of a critical point m_i is the number of linearly independent directions around m_i in which f decreases. So a minimum has index 0, a maximum has index d, and a saddle point has index between 1 and $d-1$. Let $M_a = f^{-1}\big((-\infty, a]\big)$ be the sublevel set corresponding to $a \in \mathbb{R}$. Morse theory tells us that each M_{a_i} is homotopy equivalent to a CW complex with one λ_i-cell for each critical point m_i. In particular, M_{a_i} is homotopy equivalent to $M_{a_{i-1}}$

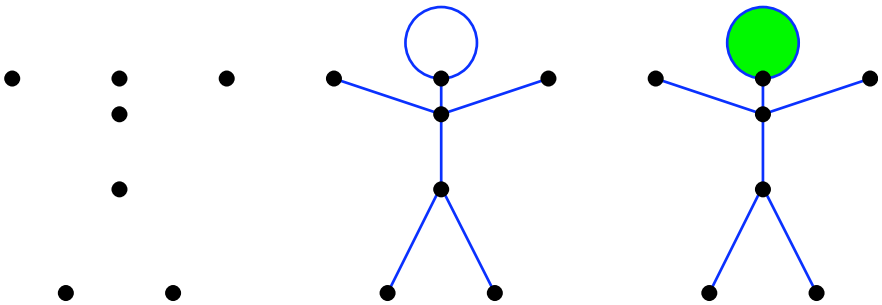

FIGURE 3. A stick figure represented as a CW complex containing seven 0-cells, seven 1-cells, and one 2-cell. The 0-skeleton is on the left, the 1-skeleton is in the center, and the full 2-skeleton is on the right.

with a single λ_i-cell attached. For instance, M_{a_1} is homotopy equivalent to a point and is obtained from M_{a_0}, the emptyset, by attaching a single 0-cell. Figure 4 contains an example.

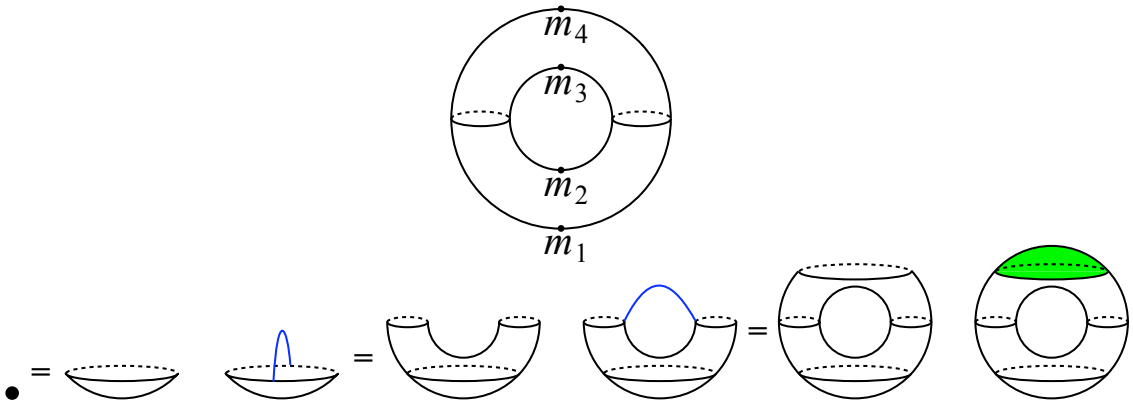

FIGURE 4. Morse theory example. (*Top*) Manifold M is a torus and function $f\colon M \to \mathbb{R}$ is the height function. There are four critical points $m_1, \ldots, m_4$ with indices 0, 1, 1, and 2 respectively. Let $a_0 < f(m_1) < a_1 \cdots < f(m_4) < a_4$. (*Bottom*) Moving from left to right, we attach cells of dimension 0, 1, 1, and 2 in order to obtain M_{a_i} from $M_{a_{i-1}}$.

Though Morse theory is traditionally stated in terms of sublevel sets, there is an equivalent formulation in terms of superlevel sets. The superlevel sets of f correspond to the sublevel sets of $-f$, so m_i is a critical point of f with index λ_i if and only if m_i is a critical point of $-f$ with index $d - \lambda_i$. It follows that superlevel set $M^{a_{i-1}} = f^{-1}\big([a_{i-1}, \infty)\big)$ is obtained from M^{a_i} by attaching a cell of dimension $d - \lambda_i$.

3.3. **Nudged elastic band.** To find saddle points we use the nudged elastic band method (NEB) from computational chemistry [Jónsson et al. 1998; Henkelman and Jónsson 2000]. Consider a chemical system, e.g. several molecules, whose space of states is parametrized by $\mathbb{R}^n$ and is equipped with a differentiable map $E\colon \mathbb{R}^n \to \mathbb{R}$ encoding the potential energy of the system at each state. The local minima of E correspond to stable states, and chemists are interested in finding reaction paths between two stable states. A reaction path is a minimum energy path, whose points minimize E in all directions perpendicular to the path, and which

passes through at least one saddle point of index one. In NEB, a path is approximated as a piecewise linear path, which we call a band. Forces move the band towards a minimum energy path. A gradient force moves each node in the band along the component of $-\nabla E$ perpendicular to the band, and a spring force prevents the band from tearing apart. Extra forces can be added to smoothen or dampen the motion.

4. Method

We suppose data set $X \subset \mathbb{R}^n$ is a finite sampling from an unknown probability density function $g\colon \mathbb{R}^n \to [0,\infty)$. We are interested in regions of high density, for if $g(x)$ is very small for some $x \in X$ then x is regarded as a noisy datapoint not representing the main features. Therefore, we would like to understand the superlevel sets

$$Y^a = g^{-1}\big([a,\infty)\big) = \{y \in \mathbb{R}^n \mid g(y) \geq a\}$$

containing all points in $\mathbb{R}^n$ with density at least $a \in [0,\infty)$. The superlevel sets Y^a encode how the dense regions of data set X are organized.

We follow the ideas of Morse theory to build CW complex models Z^a approximating the superlevel sets Y^a. First we use X to build a differentiable density estimate $f\colon \mathbb{R}^n \to [0,\infty)$ approximating g. The 0-cells of our models will be local maxima of f. The 1-cells will be paths of high density between 0-cells, found using NEB. We suggest an adaptation of NEB to find dense 2-cells with boundaries in the 1-skeleton, and analogously for higher dimensions. Given a cell $e \subset \mathbb{R}^n$, let its density be $\inf\{f(y) \mid y \in e\}$. We define Z^a to be the union of the cells with density greater than or equal to a.

Note that $Z^a \subset Z^b$ for $a \geq b$. A feature of our approach is that we produce the CW complex model Z^a for many values of a at once. This allows the user to observe how Z^a grows as a decreases. In accordance with the idea of topological persistence [Edelsbrunner et al. 2002; Zomorodian and Carlsson 2005; Edelsbrunner and Harer 2010], knowing Z^a over a range of values for a gives a better picture of the data set than knowing Z^a for any single choice of a. For example, in Section 5.1 we produce three different models—four vertices, a square loop, and a square disk—for a social network data set (Figure 6). Each model represents the data at a different density threshold, and together they provide a more complete understanding than any single CW complex model. In addition, our nested output can be used to measure the significance of topological features. Suppose a 1-cell with density a appears in Z^a to form a new loop, and suppose a 2-cell with density $b \leq a$ appears in Z^b to fill this loop to a disk. Then $a - b$ measures how long this loop persists. If $a - b$ is large then this loop is likely a significant feature of the data set, but if $a - b$ is small then the loop may be the product of noise.

Our method depends on the choice of several parameters. The main parameter which we find necessary to tune is the standard deviation used to build a density estimator. We use a consistent choice for all other parameters across all of our test data sets, which suggests that the other parameters are not as difficult to select.

Several steps in our approach can be treated in a variety of ways. To name a few, we estimate density, find local maxima, generate random initial bands, simulate bands according to a formulation of NEB, and cluster. We use simple methods to handle each of these steps, but we leave open the possibility of substituting more sophisticated methods.

4.1. **Density estimation.** Using X, we build the differentiable density estimator $f\colon \mathbb{R}^n \to [0,\infty)$ approximating g as follows. Let $\psi_{x,\sigma}\colon \mathbb{R}^n \to [0,\infty)$ be the probability density of a normal distribution centered at $x \in \mathbb{R}^n$ with standard deviation $\sigma > 0$. More explicitly, the $n \times n$ covariance matrix contains σ^2 along the diagonal and 0 elsewhere. We use the kernel estimator $f(y) = |X|^{-1} \sum_{x\in X} \psi_{x,\sigma}(y)$. See Silverman [1986] for other possible estimators, including different kernels or adaptive choices of standard deviation.

We regard the selection of σ as the most important parameter choice. One may increase σ to smoothen f or decrease σ to expose local detail.

4.2. **0-cells.** To find 0-cells, we pick a random sample of points from X and flow each along its gradient towards a local maxima of f using the mean shift iterative procedure [Fukunaga and Hostetler 1975; Cheng 1995]. Let $y_0 \in X$ be one of the initial points. We define a sequence $y_0, y_1, \ldots$ by setting $y_{i+1} = m(y_i)$, where the mean shift function $m\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$m(y) = \frac{\sum_{x\in X} \psi_{x,\sigma}(y)x}{\sum_{x\in X} \psi_{x,\sigma}(y)} \ .$$

The vector $m(y) - y$ is proportional to the normalized gradient $\nabla f(y)/f(y)$, and the y_i converge to a local maxima of f.

It is necessary to cluster the convergent points to identify which represent the same 0-cell, and we select the densest member from each cluster as a 0-cell. We make the following consistent parameter choices with all test data sets. We say point y_i has converged when $\|m(y_i) - y_i\|$ is less than 10^{-4}, and we perform single-linkage clustering with threshold distance 0.3.

4.3. **1-cells.** To find 1-cells we use NEB, which we now describe in our data analysis setting. Our formulation is similar to Jónsson et al. [1998] and Henkelman and Jónsson [2000]. A piecewise linear band is given by a list of nodes $v_1, v_2, \ldots, v_N$, with endpoints v_1 and v_N in our set of 0-cells. Forces act on the intermediate nodes v_i with $1 < i < N$ while the endpoints remain fixed. The first task is to approximate a unit tangent vector τ_i at each intermediate node. Define $u_i^+ = v_{i+1} - v_i$ and $u_i^- = v_i - v_{i-1}$. We use a naïve tangent estimate $\tau_i = (u_i^+ + u_i^-)/\|u_i^+ + u_i^-\|$ given by averaging. Henkelman and Jónsson [2000] use a more elaborate tangent.

At each intermediate node v_i we define a total force

$$F_i \ = \ c\nabla f(v_i)|_\perp \ + \ (\|u_i^+\| - \|u_i^-\|)\tau_i \ + \ \mathit{smoothing}. \tag{1}$$

The expression $\nabla f(v_i)|_\perp$ is the component of $\nabla f(v_i)$ perpendicular to the tangent τ_i, and is called the gradient force. The gradient constant c adjusts the strength of the gradient force. To normalize with respect to the maximum gradient of the normal distribution, we set

$$c = \big(\sup_{y\in\mathbb{R}^n} \|\nabla\psi_{\vec{0},\sigma}(y)\|\big)^{-1} = \big(\sigma\sqrt{2\pi}\big)^n\sqrt{e}.$$

The term $(\|u_i^+\| - \|u_i^-\|)\tau_i$ is the spring force. This name is slightly misleading, as the spring force neither enforces a natural spring length on each edge nor minimizes the length of each edge. Instead, the spring force aims to equate the lengths of adjacent edges in the band.

In Eq. [1], *smoothing* stands for a smoothing force added to prevent kinks, which hinder convergence, from forming in the band. Let θ_i be the angle between vectors u_i^+ and u_i^-. Typically these vectors are close to parallel and θ_i is close to zero. Let $0 \le \alpha < \beta \le \pi$

be fixed angles, and let $h_{\alpha,\beta}\colon [0,\pi] \to [0,1]$ be zero for $x \le \alpha$, one for $x \ge \beta$, and increase continuously from zero to one as x increases from α to β. For instance, we set

$$h_{\alpha,\beta}(x) = \begin{cases} 0 & \text{if } x \le \alpha \\ \big(1 - \cos(\pi \frac{x-\alpha}{\beta-\alpha})\big)/2 & \text{if } \alpha < x < \beta \\ 1 & \text{if } x \ge \beta. \end{cases}$$

We define our smoothing force to be $h_{\alpha,\beta}(\theta_i)(u_i^+ - u_i^-)$, which moves v_i in order to decrease θ_i whenever $\theta_i > \alpha$.

Evolving the band amounts to numerically solving the system of first order differential equations $v_i' = F_i$. Figure 16 in Appendix B shows the motion of a sample band. Note we use the first derivative v_i' rather than the second derivative. In the chemistry setting, it is appropriate to set acceleration proportional to the gradient of the potential energy. In our setting, we do not view ∇f as a force in the literal sense but only as an indication of which direction to move in order to maximize f. As the first order equation is simple and gives good results, we use it. Nevertheless, we have also had success with the second order equation and do not dismiss its use.

Let p and q be distinct 0-cells. To find the 1-cells between p and q we generate a sample of initial bands joining them, as described in Appendix B. We evolve each band until it converges and discard non-convergent bands. We also discard convergent bands which pass too close to any 0-cell $r \ne p, q$, as such a band should instead appear as the concatenation of two others.

To identify which bands represent the same 1-cell between p and q we cluster the bands. We define a metric $d_{p,q,N}$ on bands of N nodes starting at p and ending at q. If $v_1, \dots, v_N$ and $\tilde{v}_1, \dots, \tilde{v}_N$ are two such bands, then let $d_{p,q,N}(\{v_i\}, \{\tilde{v}_i\}) = (N-2)^{-1} \sum_{i=2}^{N-1} d(v_i, \tilde{v}_i)$. From each cluster we select the band with the highest density to represent the 1-cell. We estimate a band's density as $\min\{f(v_1), \dots, f(v_N)\}$, though one could obtain a more accurate estimate by subdividing the band further or by using the climbing image method of Henkelman et al. [2000].

We make the following consistent parameter choices for all of the test data sets. We include $N = 11$ nodes in the bands and let the smoothing force angle parameters be $\alpha = \pi/6$ and $\beta = \pi/2$. We set gradient constant c as above and say a band has converged when $(N-2)^{-1} \sum_{i=2}^{N-1} \|v_i'\|$ is less than 10^{-4}. If a convergent band is within distance 0.5 of another 0-cell, we discard the band. We perform single-linkage clustering with threshold distance 0.3.

4.4. **Higher-dimensional cells.** In order to search for higher-dimensional cells, one can imagine adapting NEB. We discuss one approach, which we use with the test data sets, in Appendix C. We also discuss several weaknesses of this approach.

5. Results

We test our method on five data sets whose sizes and dimensions are shown in Table 1. Please see Appendix A for further information about the data sets. We have implemented our method with Java code, which is available along with four of the data sets at the following webpage: `http://code.google.com/p/neb-tda`.

TABLE 1. Data set information

	social network	optical	range	optical flow	gene
size of data set X	1,127	15,000	15,000	15,000	43
dimension n	5	8	24	16	30
standard deviation[a] σ	0.45	0.20	0.35	0.30	1.35

[a] This is not a property of the data set, but instead the standard deviation we use to estimate density.

5.1. **Social network.** The National Longitudinal Study of Adolescent Health is a school-based study of American youth. In 1994-95 a sample of high schools and middle schools was chosen, and when possible each high school was paired with a sister middle school from the same community. The students from each school were asked to list up to five of their closest female friends who attend their school or their sister school, and likewise for their male friends.

Moody [2001] creates network graphs from the survey data. Each student is represented by a vertex, and the edge between two students exists if each student listed the other as a close friend. We analyze the graph for the students from "Countryside High School," a pseudonym used by Moody, and its sister middle school. Of the 1,147 students from this community who participated in the survey, 20 shared no connections with any others and were removed from the graph.

Figure 5 illustrates the following interesting structure in this graph. Most of the students can be placed into one of four groups, containing a majority of white high schoolers, non-white high schoolers, white middle schoolers, or nonwhite middle schoolers. There are many friendships between students in the same group. In addition, there are a significant number of friendships between groups of the same school or race category. However, their are very few friendships between groups with neither category in common.

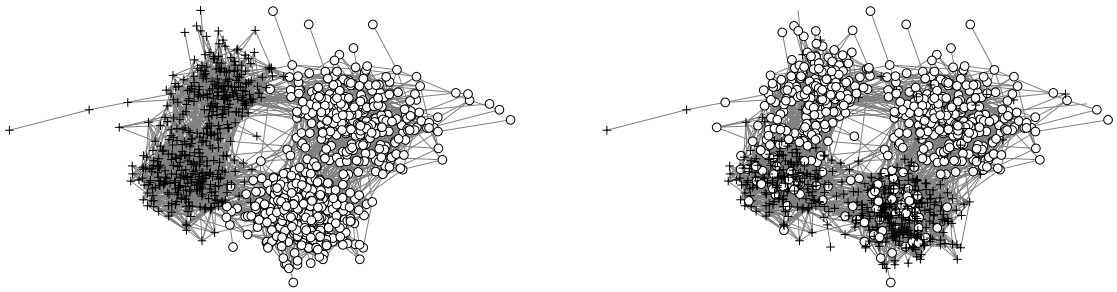

FIGURE 5. Social network for "Countryside High School" and its sister middle school. (*Left*) Cross vertices are students from the middle school and circle vertices are students from the high school. (*Right*) Cross vertices are white students and circle vertices are nonwhite students. A handful of students are without race data and their vertices are left unmarked.

Our goal is to use our Morse theory approach to recover this structure. We use stress majorization, an optimization strategy in multidimensional scaling [Cox and Cox 2001], to

embed the vertices of the graph as a data set $X \subset \mathbb{R}^5$ in a manner distorting the shortest path metric as little as possible. The shortest path distance between vertices v and w is the fewest number of edges one must cross to travel from v to w. Each point $x \in X$ corresponds to a student, and we regard the resulting set X as our input data set.

Our method finds four 0-cells, which correspond to the four groups of students. We find four 1-cells which form a square loop. The 0-cells have densities in $[1.69 \cdot 10^{-2},\ 2.02 \cdot 10^{-2}]$ and the 1-cells in $[1.07 \cdot 10^{-2},\ 1.54 \cdot 10^{-2}]$.[1] We find a 2-cell filling the loop with density $0.79 \cdot 10^{-2}$.

Recall CW complex Z^a is defined to be the union of the cells with density greater than or equal to a. For $a \in (1.54 \cdot 10^{-2},\ 1.69 \cdot 10^{-2})$ the model Z^a consists of four 0-cells (Figure 6). Hence we recover the groupings of students based on school and race. For $a \in (0.79 \cdot 10^{-2},\ 1.07 \cdot 10^{-2})$ the model Z^a is a square (Figure 6). The square reveals that groups sharing school or race category are more closely linked than groups sharing neither. This suggests that it is more difficult to cross two cultural barriers than one. For $a < 0.79 \cdot 10^{-2}$ the model Z^a fills to a disk (Figure 6), which is an appropriate representation of the data at a sufficiently coarse scale. Note that this increasing sequence of CW complex models provides a better understanding of the data set than any single model alone.

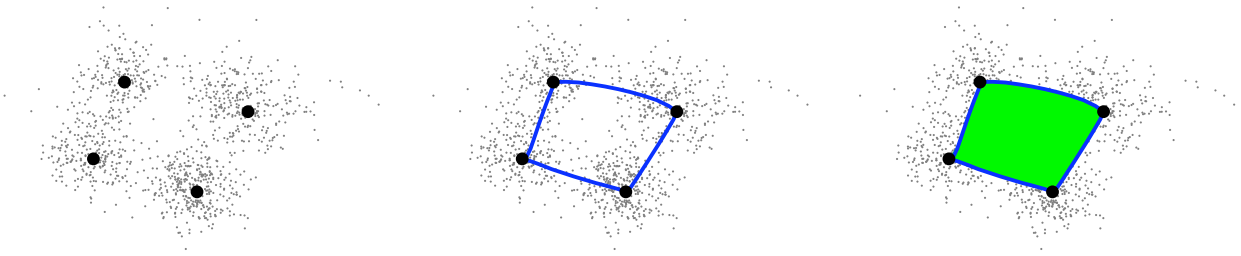

FIGURE 6. Social network data, projected to a plane using principal component analysis (PCA). Complex Z^a grows as density threshold a decreases. From left to right, we have four 0-cells, a square loop, and a disk.

5.2. **Optical image patches.** The optical image database collected by van Hateren and van der Schaaf [1998] contains a variety of indoor and outdoor scenes. From this database, Lee et al. [2003] select a large random sample of 3×3 patches. Each patch is thought of as a point in $\mathbb{R}^9$ with coordinates the logarithms of grayscale pixel values. Lee et al. define a norm measuring the contrast of a patch and select the high-contrast patches. They normalize each patch by subtracting the average patch value and dividing by the contrast norm. Lastly, Lee et al. change to the Discrete Cosine Transform (DCT) basis $\{e_1, e_2, \ldots, e_8\}$, which maps the patches to the unit sphere $S^7 \subset \mathbb{R}^8$. Note e_1 and e_2 are horizontal and vertical linear gradients and e_3 and e_4 are horizontal and vertical quadratic gradients (Figure 12 in Appendex A). Let $\mathcal{M}$ be the resulting set of high-contrast, normalized, 3×3 patches. Though $\mathcal{M}$ fills out all of S^7, some regions of the sphere are more dense than others.

Carlsson et al. [2008] use persistent homology to study $\mathcal{M}$. Using a family of density estimators they select a family of dense core subsets from $\mathcal{M}$. They apply persistent homology, and for a global estimate of density their core subset has $\operatorname{rank}(H_0) = \operatorname{rank}(H_1) = 1$,

[1] Occasionally we also find a 0-cell with density below $6 \cdot 10^{-4}$, but due to its low density value it does not affect our analysis.

where H_i denotes the i-th homology group. This core subset lies near the primary circle $\{\alpha e_1 + \beta e_2 \mid (\alpha, \beta) \in S^1\}$ containing linear gradients at all angles (Figure 7). With a more local estimate of density, the core subset has $\text{rank}(H_0) = 1$ and $\text{rank}(H_1) = 5$. Carlsson et al. identify a three circle model matching this homology profile and the data. In addition to the primary circle, the three circle model contains two secondary circles, $\{\alpha e_1 + \beta e_3 \mid (\alpha, \beta) \in S^1\}$ and $\{\alpha e_2 + \beta e_4 \mid (\alpha, \beta) \in S^1\}$, which include quadratic gradients in the horizontal or vertical direction (Figure 7). The primary circle reflects nature's preference for linear gradients in all directions and the secondary circles reflect nature's preference for the horizontal and vertical directions.

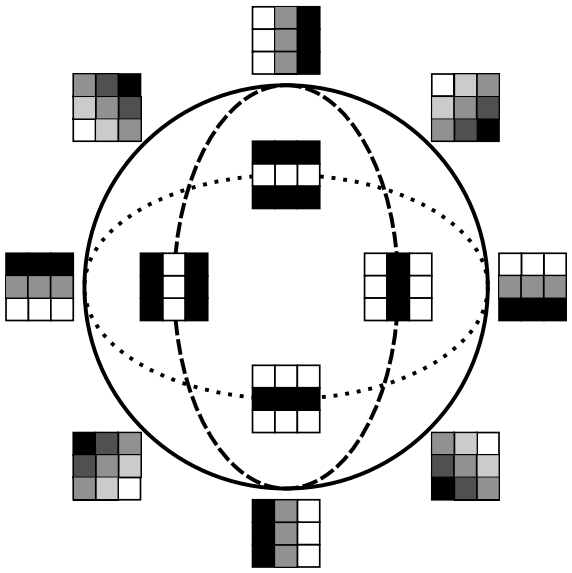

FIGURE 7. Three circle model. The solid outer circle is the primary circle and contains linear gradients. The dotted and dashed inner circles are the horizontal and vertical secondary circles which contain quadratic gradients. Each secondary circle intersects the primary circle twice, but the secondary circles do not intersect each other.

For this paper we select a random input data set $X \subset \mathcal{M}$ of size 15,000. We find four 0-cells located near the four most common patches $\pm e_1$ and $\pm e_2$. Between each of the four 0-cell pairs $\{e_1, e_2\}$, $\{e_2, -e_1\}$, $\{-e_1, -e_2\}$, and $\{-e_2, e_1\}$ we find a quarter-circular 1-cell. Together these form the primary circle. Between the pair $\{e_1, -e_1\}$ we find two semicircular 1-cells forming the horizontal secondary circle, and between $\{e_2, -e_2\}$ we find two semicircular 1-cells forming the vertical secondary circle. The 0-cells have densities in $[2.11, 2.36]$,[2] the primary circle 1-cells in $[1.17, 1.28]$, and the secondary circle 1-cells in $[0.33, 0.38]$. For $a \in (1.28, 2.11)$ our model Z^a is the four most common patches, for $a \in (0.38, 1.17)$ it is the primary circle, and for $a < 0.33$ it is the three circle model (Figure 8).

Carlsson et al. [2008] discover a 2-dimensional Klein bottle surface that contains the three circle model as its backbone and that fits a core subset of $\mathcal{M}$. The Klein bottle model is a good example of how data set models can be used not only to improve understanding but

[2]Occasionally we also find a 0-cell with density below 0.08, but due to its low density value it does not affect our analysis.

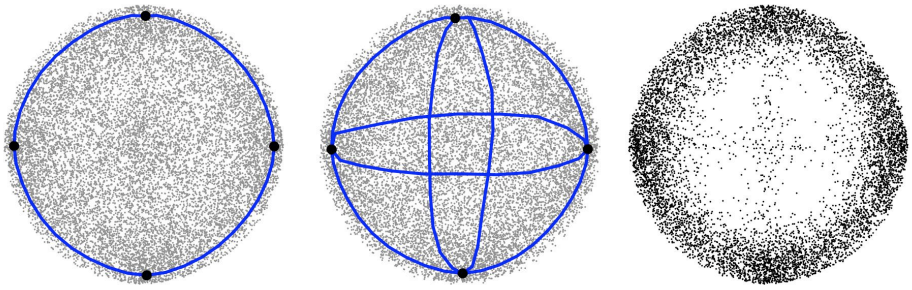

FIGURE 8. (*Left*) Optical image patches X and primary circle complex Z^a for $a \in (0.38, 1.17)$, projected to the e_1e_2 plane. (*Center*) Three circle model Z^a for $a < 0.33$, projected to the plane spanned by $e_1 + \frac{1}{4}e_4$ and $e_2 + \frac{1}{4}e_3$ so that the secondary circles are visible. (*Right*) The densest 8,500 points of X, projected to the e_1e_2 plane. The primary circle appears clearly and the projections of the secondary circle patches form a faint cross.

also to improve applications. As a low-dimensional manifold, the Klein bottle can be used in image compression schemes [Carlsson et al. 2008]. In addition, the Klein bottle model is being used to identify and analyze optical image textures [Perea and Carlsson].

5.3. **Range image patches.** In a range image each pixel stores the distance between the laser scanner and the nearest object in the corresponding direction. Lee et al. [2003] select a large random sample of log-valued, high-contrast, normalized, 3×3 range image patches from the Brown database, which contains a variety of indoor and outdoor scenes. They observe that the patches cluster near binary patches, where the binary values correspond to foreground and background.

Though the largest clusters are arranged in the shape of a circle, the 3×3 binary patches are too coarse to fill the full circle. Adams and Carlsson [2009] consider 5×5 patches, preprocessed in manner similar to Lee et al. [2003]. They obtain a large sample $\mathcal{R}$ of high-contrast, normalized, 5×5 range image patches. The 5×5 DCT basis now contains 24 vectors $\{e_1, e_2, \ldots, e_{24}\}$, where e_1 and e_5 are horizontal and vertical linear gradients (Figure 13 in Appendix A). With persistent homology they find that a dense core subset of $\mathcal{R}$ is well modeled by the range primary circle $\{\alpha e_1 + \beta e_5 \mid (\alpha, \beta) \in S^1\}$.

For our Morse theory approach, we pick a random subset $X \subset \mathcal{R}$ of size 15,000. We find four 0-cells near $\pm e_1$ and $\pm e_5$. Three of these 0-cells have densities in $[0.59, 0.75]$ while the 0-cell near e_1 has density 1.47. This reflects the fact that many range patches are shots of the ground and are hence near the horizontal linear gradient given by e_1. We find four 1-cells forming a loop with densities in $[0.51, 0.64]$, and a 2-cell filling the loop with density 0.39. For $a \in (0.75, 1.47)$ our model Z^a is the single 0-cell near e_1, and for $a \in (0.39, 0.51)$ our model Z^a is the range primary circle (Figure 9).

Any method in data analysis must tolerate some amount of noise, that is points $x \in X$ with low density. One may try to remove noisy points from the data set, but in general it is not easy to remove noise without also removing features. Because we use data set X primarily to build density estimate $f \colon \mathbb{R}^n \to \mathbb{R}$, we believe that noise which is sparse enough to not significantly affect f will not significantly affect our CW complex output. As evidence for this claim, consider the large samples $\mathcal{M}$, $\mathcal{R}$ of optical and range image patches. Carlsson

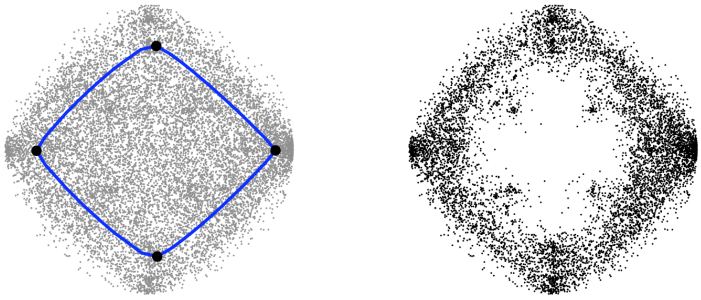

FIGURE 9. (*Left*) Range image patches X and range primary circle Z^a for $a \in (0.39, 0.51)$, projected to the e_1e_5 plane. (*Right*) The densest 9,000 points of X.

et al. [2008] and Adams and Carlsson [2009] remove noise from $\mathcal{M}$ and $\mathcal{R}$ by studying only core subsets of points with densities in the top 30%. However, in this paper we work with random subsets of $\mathcal{M}$ and $\mathcal{R}$ containing both dense and noisy points.

5.4. **Optical flow patches.** A video records a sequence of images, and the apparent motion in the images is called optical flow. At each frame, the optical flow is represented by a vector field with one vector per pixel that points to where that pixel appears to move for the subsequent frame. No instrument measures optical flow directly, and estimating optical flow from a video is an ill-posed problem. However, Roth and Black [2007] create a database of ground-truth optical flow by pairing range images with camera motions and calculating the produced optical flow. The range images are from the Brown database, and the camera motions are retrieved from a database of videos from hand-held or car-mounted cameras.

With preprocessing steps analogous to Lee et al. [2003] we create a large sample $\mathcal{F}$ of high-contrast, normalized, 3×3 optical flow patches. We change to the DCT basis $\{e_1^u, \ldots, e_8^u, e_1^v, \ldots, e_8^v\}$, where the superscript u denotes flow in the horizontal direction and v denotes the vertical direction (Figure 14 in Appendix A). We select X to be a random subset of $\mathcal{F}$ of size 15,000.

We find four 0-cells near $\pm e_1^u$ and $\pm e_2^u$ and four 1-cells which form a loop. These cells have densities in $[0.71, 2.79]$. We find a 2-cell filling the loop with density 0.39. For $a \in (0.39, 0.71)$ our model Z^a recovers the horizontal flow circle near $\{\alpha e_1^u + \beta e_2^u \mid (\alpha, \beta) \in S^1\}$ (Figure 10). Note that the horizontal flow circle is obtained by applying horizontal camera motion to range patches from the range primary circle. Also, one expects horizontal camera motion to be more common than vertical motion in hand-held and car-mounted videos. Therefore, the horizontal flow circle combines important patterns from both the range image and camera motion databases.

5.5. **Gene expression data.** Spellman et al. [1998], Alter et al. [2000], and Whitfield et al. [2002] identify genes associated with the cell cycle using microarray analysis, and the expression levels of these genes are cyclical. One way to model cyclic behavior mathematically is with a periodic function. However, while a periodic function has a fixed amplitude and period, cyclic behavior in biology is not as rigid. The amplitude and period of cell cycle gene expressions change with time. In particular the amplitude tends to decrease, which

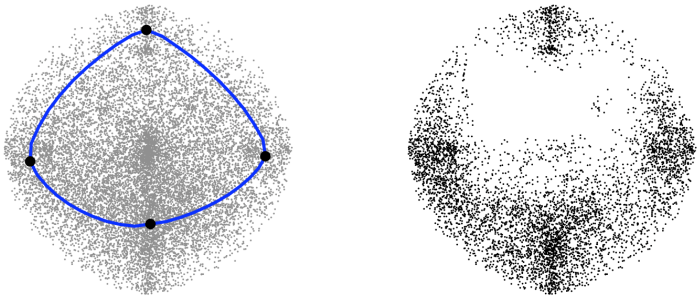

FIGURE 10. (*Left*) Optical flow patches X and horizontal flow circle Z^a for $a \in (0.39, 0.71)$, projected to the $e_1^u e_2^u$ plane. (*Right*) The densest 6,000 points of X.

can be interpreted as the dephasing of initially synchronized cells. Another way to model cyclic behavior mathematically is with a circle, and this topological representation is robust to changes in amplitude and frequency.

Whitfield et al. [2002] measure gene expression levels as HeLa cells proceed through the cell cycle. We select genes whose measurements pass several quality thresholds and which deviate significantly from the average expression. Our thresholding is stringent as the purpose of this exercise is not to discover new biology but to test our method on time series data. We normalize and cluster the resulting genes. One cluster consists of six genes which are well-known to be part of the cell cycle and which have periodic expression levels, as shown in Figure 15 in Appendix A. For time point $t \in \{0, 1, \ldots, 46\}$, let w_t be the vertical vector of length six containing the expression levels of these genes at time t. We use blocks of five time points to form our data set. We set $X = \{x_0, \ldots, x_{42}\} \subset \mathbb{R}^{30}$, where

$$x_0 = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_4 \end{pmatrix}, \quad x_1 = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_5 \end{pmatrix}, \quad \ldots, \quad x_{42} = \begin{pmatrix} w_{42} \\ w_{43} \\ \vdots \\ w_{46} \end{pmatrix}.$$

We apply our method to X and find three 0-cells and three 1-cells which form a circle. These cells have densities in $[9.4 \cdot 10^{-18},\ 12.2 \cdot 10^{-18}]$. We find a 2-cell filling the circle with density $4.0 \cdot 10^{-18}$. The model Z^a for $a \in (4.0 \cdot 10^{-18},\ 9.4 \cdot 10^{-18})$ is a circle which recovers the cyclical nature of the gene expressions (Figure 11).

6. Conclusion

We have introduced a method for finding structure in high-dimensional Euclidean data sets. Following Morse theory, we use the nudged elastic band method to sample cells from the analogue of the Morse complex determined by the density function. We produce a nested family of CW complex models representing the dense regions of the data. We test the approach on a variety of data sets and find compact complexes revealing important nonlinear patterns, suggesting that the method may be a helpful tool for understanding new data sets.

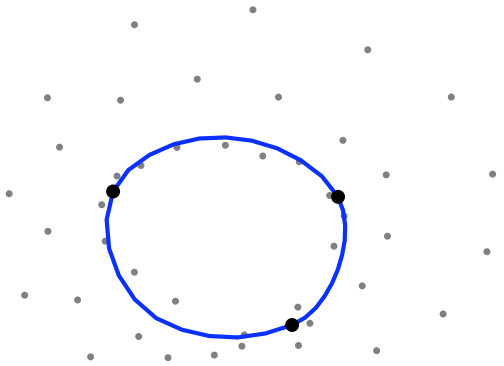

FIGURE 11. Gene expression data set X, projected to a plane using PCA. Complex Z^a is a circle for $a \in (4.0 \cdot 10^{-18},\ 9.4 \cdot 10^{-18})$.

Appendices

APPENDIX A. ADDITIONAL DATA SET INFORMATION

We have implemented our method with Java code which is available at the following webpage: `http://code.google.com/p/neb-tda`. The optical image patches, range image patches, optical flow patches, and gene expression data sets are also available at this webpage. The raw data for the social network data set may be obtained as explained in the following section.

We provide further information about each of the test data sets.

A.1. **Social network.** More information is available at the Adolescent Health Study webpage: `http://www.cpc.unc.edu/projects/addhealth`. To obtain a copy of the non-linkable Adolescent Health Network Structure files, please contact `addhealth@unc.edu`.

A.2. **Optical image patches.** The optical image database collected by van Hateren and van der Schaaf [1998] contains 4,167 grayscale images from indoor and outdoor scenes, each 1020×1532 pixels (Figure 12). More details and the database itself are available at `http://www.kyb.mpg.de/bethge/vanhateren/index.php`. From this database, Lee et al. [2003] create a set $\mathcal{M}$ of high-contrast, normalized, 3×3 patches through the following preprocessing steps.

(1) Lee et al. select a large random sample of 3×3 patches with coordinates the logarithms of grayscale pixel values.

$$\begin{bmatrix} x_1 & x_4 & x_7 \\ x_2 & x_5 & x_7 \\ x_3 & x_6 & x_9 \end{bmatrix}$$

Each patch is represented by a vector $x = (x_1, \ldots, x_9)^T \in \mathbb{R}^9$.

(2) Lee et al. define a norm $\| \ \|_D$ measuring the contrast of a patch. Two coordinates x_i and x_j of x are neighbors, denoted $i \sim j$, if the corresponding pixels in the 3×3

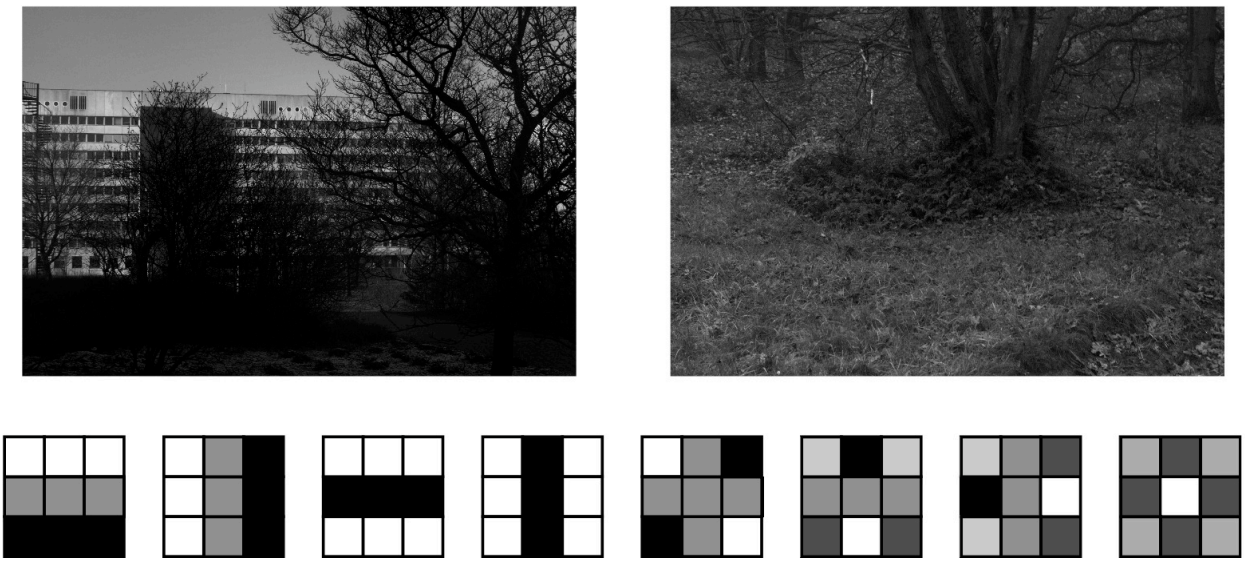

FIGURE 12. (*Top*) Sample optical images from the van Hateren database. (*Bottom*) Vectors $e_1, e_2, \ldots, e_8$ from the 3×3 DCT basis.

patch are adjacent. Let $\|x\|_D = \sqrt{\sum_{i\sim j}(x_i - x_j)^2}$. Lee et al. select the patches with contrast norm in the top 20% of their sample.

(3) Lee et al. normalize each patch by subtracting the average patch value and by dividing by the contrast norm. This maps the patches to a 7-dimensional ellipse.
(4) Lee et al. change to the Discrete Cosine Transform (DCT) basis $\{e_1, \ldots, e_8\}$ for 3×3 patches. The basis vectors are normalized to have contrast norm one, and so this maps the patches to a 7-dimensional sphere.

Let $\mathcal{M}$ be the resulting set of high-contrast, normalized, 3×3 optical patches. Carlsson et al. [2008] study dense core subsets from $\mathcal{M}$ using persistent homology. In this paper we select a random test data set $X \subset \mathcal{M}$ of size 15,000.

A.3. **Range image patches.** The Brown range image database by Lee and Huang is a set of 197 range images from indoor and outdoor scenes, mostly 444×1440 pixels (Figure 13). The operational range for the Brown scanner is typically 2-200 meters, and the distance values for the pixels are stored in units of 0.008 meters. The database can be found at the following webpage: `http://www.dam.brown.edu/ptg/brid/index.html`.

From the Brown database we obtain a space of range image patches through the following steps, which are similar to the procedures of Lee et al. [2003] and Adams and Carlsson [2009].

(1) We randomly select about $4 \cdot 10^5$ size 5×5 patches from the images in the database. Each patch is represented by a vector $x \in \mathbb{R}^{25}$ with logarithm values.
(2) We compute the contrast norm $\|x\|_D = \sqrt{\sum_{i\sim j}(x_i - x_j)^2}$ of each patch and select the patches with contrast norm in the top 20% of the entire sample.
(3) We subtract from each patch the average of its coordinates and divide by the contrast norm.

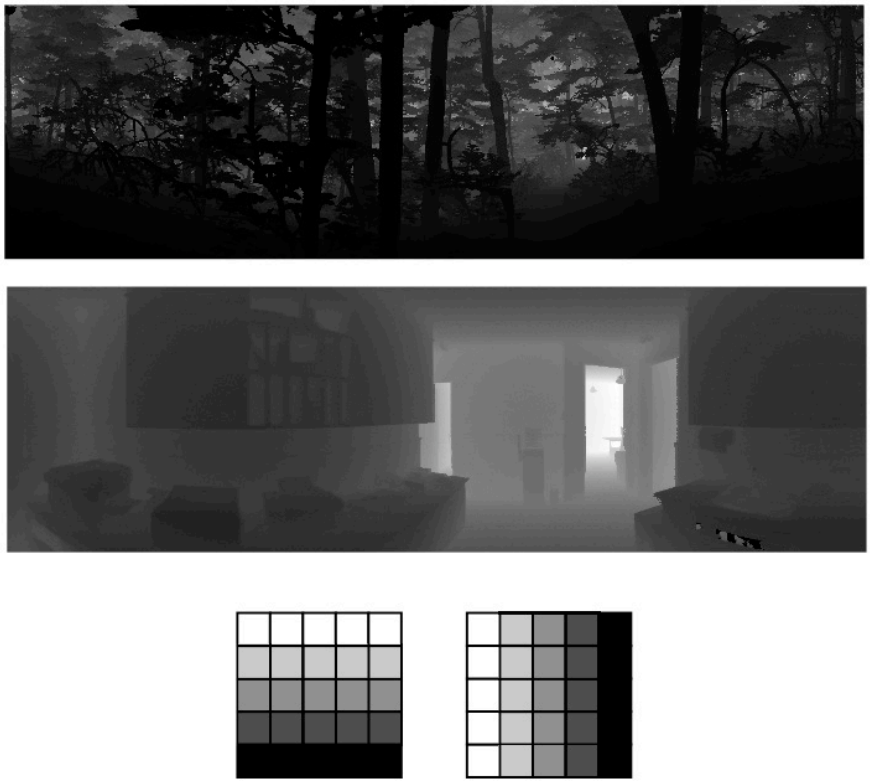

FIGURE 13. (*Top*) Sample range images from the Brown database. (*Bottom*) Vectors e_1 and e_5 from the 5×5 DCT basis are horizontal and vertical linear gradients.

(4) We change to the DCT basis $\{e_1, \ldots, e_{24}\}$ for 5×5 patches, normalized to have contrast norm one. This maps the patches to a 23-dimensional sphere.

Let $\mathcal{R}$ be the resulting set of high-contrast, normalized, 5×5 range patches. Our test data set is a random subset $X \subset \mathcal{R}$ of size 15,000.

A.4. **Optical flow patches.** The optical flow database by Roth and Black [2007] contains 800 optical flow fields, each 250×200 pixels (Figure 14). This ground-truth optical flow is generated by pairing range images and camera motions. The range images are from the Brown range image database. The camera motions are retrieved from a database of 67 videos from hand-held or car-mounted video cameras, each approximately 100 frames long. The camera motions are extracted using *boujou* software by 2d3 Ltd., available at `http://www.2d3.com`. The Roth and Black database is available at `http://www.gris.informatik.tu-darmstadt.de/~sroth/research/flow/downloads.html`.

From the Black and Roth database we create a space of optical flow patches, and our preprocessing is similar to that of Lee et al. [2003].

(1) We randomly choose $4 \cdot 10^5$ size 3×3 optical flow patches from the Roth and Black database. Each patch is a matrix of ordered pairs, where u_i and v_i are the horizontal

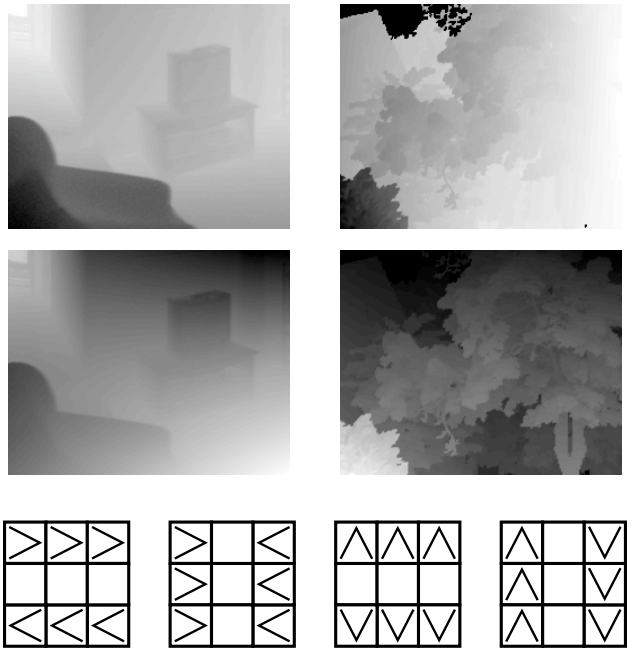

FIGURE 14. (*Top*) Two sample optical flows from the Roth and Black database. Horizontal components are on the top and vertical components are on the bottom. White corresponds to flow in the positive direction ($+x$ or $+y$) and black corresponds to the negative direction. (*Bottom*) Vector fields e_1^u, e_2^u, e_1^v, and e_2^v from the 3×3 optical flow DCT basis. Compare with the patches e_1 and e_2 in Figure 12.

and vertical components, respectively, of the flow vector at pixel i.

$$\begin{bmatrix} (u_1, v_1) & (u_4, v_4) & (u_7, v_7) \\ (u_2, v_2) & (u_5, v_5) & (u_8, v_8) \\ (u_3, v_3) & (u_6, v_6) & (u_9, v_9) \end{bmatrix}$$

We define $u = (u_1, \ldots, u_9)^T$ and $v = (v_1, \ldots, v_9)^T$ to be the vectors of horizontal and vertical flow components. We rearrange each patch to be a vector

$$x = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^{18}.$$

(2) We compute the contrast norm $\|x\|_D = \sqrt{\sum_{i \sim j} \|(u_i, v_i) - (u_j, v_j)\|^2}$ for all patches. We select the patches with contrast norm in the top 20% of the entire sample.

(3) We normalize the patches to have zero average flow. More explicitly, given a patch x, let $\overline{u} \in \mathbb{R}^9$ have each entry equal to $\frac{1}{9} \sum_{i=1}^{9} u_i$, the average of the horizontal components. Let $\overline{v}$ be defined similarly. We replace each patch x with

$$\begin{pmatrix} u - \overline{u} \\ v - \overline{v} \end{pmatrix}.$$

We then divide each patch by its contrast norm.

(4) We change to the DCT basis $\{e_1^u, \ldots, e_8^u, e_1^v, \ldots, e_8^v\} \subset \mathbb{R}^{18}$ for 3×3 optical flow patches, where

$$e_i^u = \begin{pmatrix} e_i \\ \vec{0} \end{pmatrix} \quad \text{and} \quad e_i^v = \begin{pmatrix} \vec{0} \\ e_i \end{pmatrix}.$$

This maps the patches to a 16-dimensional sphere.

Let $\mathcal{F}$ be the resulting set of high-contrast, normalized, 3×3 optical flow patches. Our test data set is a random subset $X \subset \mathcal{F}$ of size 15,000.

A.5. **Gene expression data.** Whitfield et al. [2002] run five parallel time series experiments measuring gene expression levels as HeLa cells proceed through the cell cycle. Their experiments are available at `http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=106&23706`. We select a small set of high quality cyclically expressed genes through the following steps. Our thresholding is very strict as the purpose of this exercise is not to discover new biology but to test our method on time series data.

(1) For quality control, we include only the spots in which either
   - channel 1 mean intensity / median background intensity > 1.5,
   - channel 2 mean intensity / median background intensity > 1.5, or
   - the spot regression correlation between the channels > 0.5.

(2) We consider only genes which appear in all five time series experiments, which contain at least 60% of the measurements in each experiment, and which contain at least 70% of the measurements in total.

(3) We now focus on the experiment "Double Thymidine Block Experiment 3" which contains hourly measurements for about $4.3 \cdot 10^4$ gene elements over 46 hours, with two measurements at time zero.

(4) We KNN impute missing measurements [Troyanskaya et al. 2001] with $k = 10$ using PAM software.

(5) We mean center each row (gene) and column (time point).

(6) We zero-transform the data by subtracting the mean of the two time zero columns from all other columns.

(7) We select the genes with at least three measurements greater than 1.5 in absolute value. Only 29 genes remain.

(8) We cluster the genes using the Pearson correlation centered metric. We select one cluster of six genes—CDC2, UBE2C, TOP2A, CCNF, AURKA, and PLK1—which are cyclically expressed (Figure 15). These genes are well-known to be part of the cell cycle.

(9) We combine the two measurements at time zero into a single measurement by averaging. We now have a set of vectors $\{w_0, \ldots, w_{46}\} \subset \mathbb{R}^6$, where w_t contains the expression levels of the six genes at hour t.

(10) We use blocks of five time points to form our test data set $X = \{x_0, \ldots, x_{42}\} \subset \mathbb{R}^{30}$, where

$$x_0 = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_4 \end{pmatrix}, \quad x_1 = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_5 \end{pmatrix}, \quad \ldots, \quad x_{42} = \begin{pmatrix} w_{42} \\ w_{43} \\ \vdots \\ w_{46} \end{pmatrix}.$$

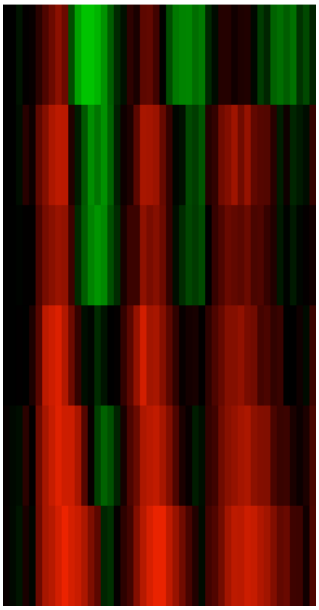

FIGURE 15. Microarray data for the six genes CDC2, UBE2C, TOP2A, CCNF, AURKA, and PLK1. Each row is a gene and each column is a time point. Red corresponds to high expression and green to low expression. Approximately three cell cycles are visible.

APPENDIX B. INITIAL BANDS

Let p and q be distinct 0-cells. Our method for generating the initial bands between p and q depends on whether data set X is a general data set in $\mathbb{R}^n$ or whether X is normalized to lie on a unit sphere $S^{n-1} \subset \mathbb{R}^n$.

For a general data set $X \subset \mathbb{R}^n$, we pick a random vector y from the set of all unit vectors perpendicular to $p - q$, which is a sphere of dimension $n - 2$. We also pick $r \in [0, d(p, q)]$ uniformly randomly. The resulting initial band is N evenly distributed nodes along the circular arc (or straight line, with probability zero) between the points p, $(p+q+ry)/2$, and q.

If X is normalized to lie on a unit sphere $S^{n-1} \subset \mathbb{R}^n$, we generate initial bands lying near S^{n-1}. Though p and q are near dense regions of X they need not lie in X nor in S^{n-1}. Let $\hat{p} = p/\|p\|$ and $\hat{q} = q/\|q\|$. We pick a random vector $y \neq \hat{p}, \hat{q}$ in S^{n-1}. The plane defined by y, $\hat{p}$, and $\hat{q}$ intersects S^{n-1} in a circle. Let $\hat{p} = \hat{v}_1, \ldots, \hat{v}_N = \hat{q}$ be the unique band that is evenly-spaced along the this intersection circle, that starts at $\hat{p}$, that ends at $\hat{q}$, and that does not pass through y. We define our initial band to be $p = v_1, \ldots, v_N = q$, where

$$v_i = \frac{(N-i)\|p\| + (i-1)\|q\|}{N-1} \, \hat{v}_i.$$

In general, one may be interested in finding dense 1-cell loops which start and end at the same 0-cell. We did not search for loops in our test data sets. See Figure 16 for a sample 1-cell trial.

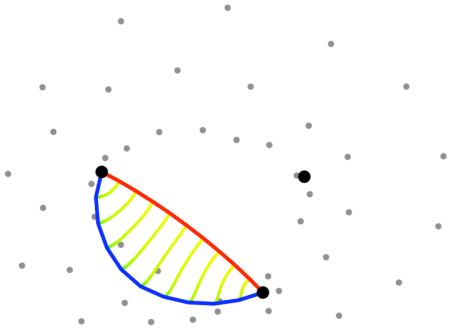

FIGURE 16. Sample 1-cell trial from the gene expression data set. The initial band is in red and the convergent 1-cell is in blue. The lines which fade from yellow to green trace the paths of the intermediate nodes. The gene expression data is projected to the plane using PCA.

APPENDIX C. HIGHER-DIMENSIONAL CELLS

One can imagine adapting the NEB method to search for higher-dimensional cells. Given an initial k-cell with $k > 1$, one would like to move it towards a maximum density k-cell. We describe a straightforward approach which we use with our test data sets. We model a k-cell as a graph $G = (V, E)$ with all edges simple. Given a node $\alpha \in V$, let $v_\alpha \in \mathbb{R}^n$ denote its position and $V_\alpha = \{\beta \in V \mid \{\alpha, \beta\} \in E\}$ denote its set of incident vertices. We estimate the k-dimensional tangent space at v_α to be the span of the first k components of a principal component analysis on $\{v_\beta - v_\alpha \mid \beta \in V_\alpha\}$. Specified boundary nodes lie in the $(k-1)$-skeleton of our CW model and remain fixed, but if α is not a boundary node, then we define

$$F_\alpha = c\nabla f(v_\alpha)|_\perp \; + \sum_{\beta \in V_\alpha} (v_\beta - v_\alpha)$$

to be the force at vertex α. As before, $\nabla f(v_\alpha)|_\perp$ is the component of $\nabla f(v_\alpha)$ perpendicular to the tangent space and is called the gradient force. Gradient constant c adjusts the strength of the force. The term $\sum_{\beta \in V_\alpha}(v_\beta - v_\alpha)$ is the spring force. We numerically solve the system of first order differential equations $v'_\alpha = F_\alpha$.

We point out several weaknesses in the straightforward approach above.

- The spring force does not generalize the dimension $k = 1$ case.
- The gradient forces and spring forces do not both go to zero; instead, they balance against one another. This means that the cell does not converge exactly to the maximum density cell. A possible remedy is to project the spring force to the tangent space. One may then need to add an appropriate smoothing force to prevent kinks from forming in the cell.
- It may be preferable to model a k-cell not as a graph with forces acting on the nodes but as a k-dimensional simplicial complex with forces acting on simplices.

- One may want an adaptive representation of a cell whose triangulation changes as the cell moves. Otherwise, the choice of an initial triangulation may affect the subsequent motion of the cell.

We leave the exploration of improved generalizations of NEB in higher dimensions as the subject of future work.

We search for 2-cells in four of our test data sets. We form a web-shaped graph with 20 nodes on each of 10 concentric rings. Let $\partial V \subset V$ denote the boundary nodes which lie in the 1-skeleton and remain fixed. We initially place the center node at the average of the boundary nodes, and we linearly interpolate between the boundary and center to place the other nodes. We set constant c as in the case of 0 and 1-cells, and we say a cell has converged when $|V \setminus \partial V|^{-1} \sum_{\alpha \in V \setminus \partial V} \|v'_\alpha\|$ is less than 10^{-3}. We estimate a convergent cell's density as $\min_{\alpha \in V} f(v_\alpha)$. See Figure 17 for the resulting 2-cells.

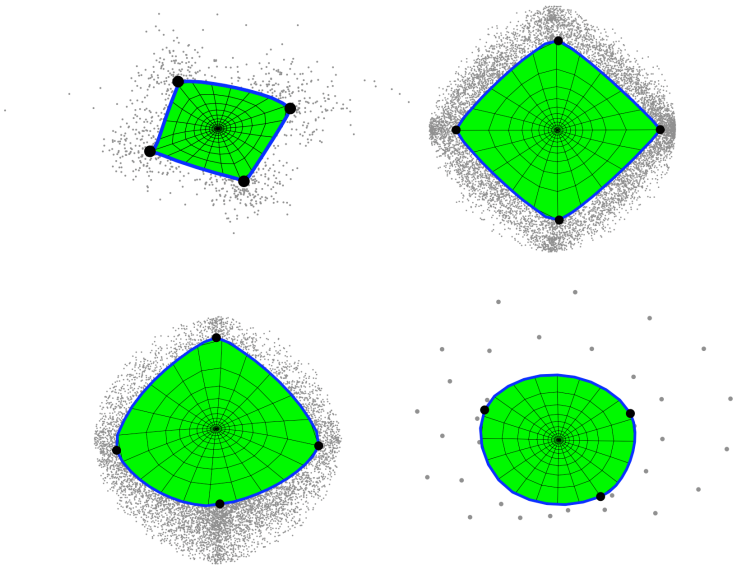

FIGURE 17. The 2-cells in the test data sets. (*Top left*) Social network data, projected to a plane using PCA. (*Top right*) Range image patches, projected to the e_1e_5 plane. (*Bottom left*) Optical flow patches, projected to the $e_1^u e_2^u$ plane. (*Bottom right*) Gene expression data, projected to a plane using PCA.

Acknowledgements. We would like to thank Monica Nicolau for her help with the gene expression data, Tim Harrington and Andrew Tausz for their help with the social network data, and Guillermo Sapiro for his help with the optical flow data. The second author would like to thank the I.I. Rabi Science Scholars Program, Columbia University. This work is supported by Office of Naval Research Grant N00014-08-1-0931, Air Force Office of Scientific Research Grant FA9550-09-1-0143, Air Force Office of Scientific Research Grant FA9550-09-0-1-0531, and National Science Foundation Grant DMS-0905823.

REFERENCES

H. Adams and G. Carlsson. On the nonlinear statistics of range image patches. *SIAM J. Imag. Sci.*, 2:110–117, 2009.

O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106, 2000.

G. Carlsson. Topology and data. *Bull. Amer. Math Soc.*, 46:255–308, 2009.

G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *Int. J. Comput. Vision*, 76:1–12, 2008.

F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. In *Proc. 27th Annu. ACM Sympos. on Comput. Geom.*, pages 97–106, June 2011.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE T. Pattern Anal.*, 17:790–799, 1995.

T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 2001.

V. de Silva and G. Carlsson. Topological estimation using witness complexes. In *Eurographics Symposium on Point-Based Graphics*, June 2004.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, 2010.

H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, 30:87–107, 2003.

R. Forman. A user's guide to discrete Morse theory. *Séminare Lotharinen de Combinatore*, 46, 2002.

K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE T. Inform. Theory*, 21:32–40, 1975.

A. Gyulassy, V. Natarajan, V. Pascucci, and B. Hamann. Efficient computation of Morse-Smale complexes for three-dimensional scalar functions. *IEEE Visualization*, 13:1440–1447, 2007.

J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

A. Hatcher. *Algebraic Topology*. Cambridge University Press, New York, 2002.

G. Henkelman and H. Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113:9978–9985, 2000.

G. Henkelman, B. P. Uberuaga, and H. Jonsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113:9901–9904, 2000.

H. Jónsson, G. Mills, and K. W. Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In B. Berne, G. Ciccotti, and D. Coker, editors, *Classical and Quantum Dynamics in Condensed Phase Systems*, pages 385–404. World Scientific, Singapore, 1998.

A. B. Lee, K. S. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *Int. J. Comput. Vision*, 54:83–103, 2003.

J. Milnor. *Morse Theory*. Princeton University Press, Princeton, 1965.

J. Moody. Race, school integration, and friendship segregation in America. *Am. J. Sociol.*, 107:679–716, 2001.

J. Perea and G. Carlsson. A Klein-bottle-based dictionary for distributions of high-contrast image patches. In preparation.

S. Roth and M. J. Black. On the spatial statistics of optical flow. *Int. J. Comput. Vision*, 74:33–50, 2007.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9: 3273–3297, 1998.

W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *J. Comput. Graph. Stat.*, 19:397–418, 2010.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.

J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, 265:359–366, 1998.

M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, 13:1977–2000, 2002.

D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In A. Cole, editor, *Numerical Taxonomy*, pages 282–311. Academic Press, London, 1969.

A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33:249–274, 2005.

Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail address: henrya@math.stanford.edu

Department of Mathematics, Harvard University, Cambridge, MA 02138
E-mail address: nasko@math.harvard.edu

Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail address: gunnar@math.stanford.edu